VideoAgent: Long-form Video Understanding with Large Language Model as Agent

Year: 2024 | *Citations: 210* | *Authors: Xiaohan Wang, Yuhui Zhang, Orr Zohar*

Abstract

Long-form video understanding represents a significant challenge within computer vision, demanding a model capable of reasoning over long multi-modal sequences. Motivated by the human cognitive process for long-form video understanding, we emphasize interactive reasoning and planning over the ability to process lengthy visual inputs. We introduce a novel agent-based system, VideoAgent, that employs a large language model as a central agent to iteratively identify and compile crucial information to answer a question, with vision-language foundation models serving as tools to translate and retrieve visual information. Evaluated on the challenging EgoSchema and NExT-QA benchmarks, VideoAgent achieves 54.1% and 71.3% zero-shot accuracy with only 8.4 and 8.2 frames used on average. These results demonstrate superior effectiveness and efficiency of our method over the current state-of-the-art methods, highlighting the potential of agent-based approaches in advancing long-form video understanding.